Reporte de Investigación

Alex Medina

30/11/2022

Indice de contenidos

Introducción

El esplendor de la era digital han traído consigo un crecimiento exponencial de los repositorios del contenido multimedia digital. En 2022, la cantidad de usuarios de teléfonos inteligentes en el mundo actual es de 6648 millones, lo que se traduce en que el 83 % de la población mundial posee un teléfono inteligente (Fuente: https://www.bankmycell.com/blog/how-many-phones-are -en el mundo), como consecuencia de esta distribución tan amplia y con la calidad cada vez mayor de las cámaras de los teléfonos inteligentes, la cantidad de fotografías tomadas en todo el mundo cada día se está disparando. En 2022, se toman 54.4 mil fotos por segundo y 1.7 billones por año.

Este desarrollo ha propiciado la creación de repositorios de imágenes bastante voluminosos de todos los campos semánticos imaginables. Estos conjuntos de datos son cada vez mas aprovechados por la industria publica y privada dentro del area de vision por computadora y aprendizaje automático para entrenar algoritmos de inteligencia artificial que que luego pueden ser aprovechados para lograr la identificación de patrones y clasificación sobre los conjuntos de images que cada vez aumentan más.

Definición del problema

Para el caso particular de esta investigación, el area de aplicación con la que se trabajará es el de prendas de vestir. Es posible resumir de manera más concreta el problema como:

> Lograr la segmentación de prendas de vestir variadas por medio de la utilización de algoritmos de Vision por Computadora y *Deep Learning*. Esto implica la creación y optimización de modelos de clasificación que serán aplicados sobre un conjunto de imágenes que contienen una prenda de vestir especifica.

El modelo será entrenado con un conjunto de imágenes de prendas de vestir variadas y etiquetas. Una vez que el modelo de entrenamiento sea lo suficientemente capaz de realizar la identificación de las prendas en las imágenes, este mismo modelo será probado sobre un conjunto nuevo.

Actualmente las soluciones para llevar a cabo la segmentación, ademas de la alternativa básica que consiste en utilizar a una persona para clasificar lo que puede observar en las imágenes, es la de la utilización de herramientas ya existentes similares de vision por computadora capaces de lograr dicha segmentación. Un ejemplo de estas herramientas es Google Cloud AutoML Vision, el cual es una herramienta de paga y en linea que puede ser utilizada incluso sin la necesidad de conocer sobre temas de desarrollo o programación para aprender como funciona.

4. Definición del problema propiamente, es decir, formular los aspectos de forma abstracta con lenguaje de matemáticas.

   (DIAGRAMA)

Para el entrenamiento de los modelos se utiliza un conjunto de datos que contienen etiquetas que identifican el tipo de prenda que está representada en la imagen a analizar, por se puede definir la investigación como un problema supervisado. Además es posible catalogarlo como un problema *online*, debido a que se pretende tener los datos completos a utilizar ya almacenados e ir mejorando la solución y los modelos con nuevos datos.

La manera en la que se estará midiendo el desempeño de los modelos es por medio de métricas básicas, se estará midiendo principalmente a través de el porcentaje de predicciones correctas que el modelo logre realizar sobre un conjunto de datos de prueba después de realizar el entrenamiento. Además también se pretende tomar en cuenta el tiempo de entrenamiento de los modelos ya que cada algoritmo puede tener un desempeño similar en cuanto a resultados pero diferente en cuanto a poder computacional requerido. Por lo tanto la selección de modelos estará basado en el equilibrio de estos dos parámetros.

Si se toma como base el trabajo que una persona podría realizar al intentar llevar a cabo esta segmentación de prendas de vestir, entonces resulta bastante simple medir el desempeño que los modelos logren, tanto por el porcentaje de predicciones correctas como por el tiempo en el que logre completar la clasificación.

Un humano con desarrollo cognitivo normal es capaz de realizar la segmentación e identificación de prendas de vestir si observa una imagen de ella con relativa facilidad. Se estima que el ojo de una persona experta en el tema llega a lograr un error de aproximadamente el 5%, esto al realizar el trabajo manual de segmentación e identificación de los objetos que se muestran sobre una imagen sobre un dataset de 1500 imágenes y 1000 clases, estos resultados son obtenidos del Large Scale Visual Recognition Challenge (ILSVRC) en el cual se ponen a prueba la experticia humana contra los modelos de vision por computadora. Por lo tanto al el objetivo es obtener al menos un 95% de predicciones correctas que representaría un desempeño similar al de una persona.

Una característica interesante de este tipo de soluciones es su escalabilidad y adaptabilidad para poder ser utilizado en problemas similares. La solución que se desarrollará debería ser posible de utilizarse con con problemas similares. Tanto la experiencia obtenida y las herramientas que se obtengan en este problema se pueden trasponer relativamente fácil sobre conjuros de datos similares.Esta es una ventaja ampliamente conocida en el area de vision por computadora, pues existen ya varios modelos e investigaciones que realizan un procedimiento similar pero aplicado sobre conjuros de datos diferentes o con contexto diferentes.

11. Lista las asunciones que has hecho hasta ahora.

    - El conjunto de datos es usable.

12. Verifica si las asunciones se cumplen.

    ☒ El conjunto de datos es usable.

Obtención de datos

1. Automatiza tanto como puedas la obtención de datos, para tener datos frescos (puede ser en script de Python/R, o Jupyter Notebook. Si son datos privados puedes omitir este paso).

   (../scripts/python/filesdownload.sh)

2. Documenta cómo obtuviste los datos y cuánto espacio ocuparán (Quarto).

   Este conjunto de datos fue obtenido desde el repositorio publico de "Kaggle". Dicho dataset ocupa un tamaño de aproximadamente 7GB de espacio en memoria. Consta de un conjunto de 5000 imágenes catalogadas en 20 tipos de prendas de vestir. Mas información sobre la obtención de los datos puede ser consultada desde el siguiente "articulo" del autor.

3. Revisa las obligaciones legales de los datos, u obtén autorización para usarlos y/o publicar derivados de ellos (escribir brevemente la situación legal en Quarto).

   El conjunto de datos Pertenece a Alexey Grigorev y fue creado como un esfuerzo de contribución comunitaria y se publica bajo una licencia Creative Commons Zero (CC0). Esto significa que cualquiera puede utilizar estos datos para cualquier fin, incluso comercial.

4. Crea un espacio de trabajo (workspace) con espacio suficiente para almacenar los datos (opcional).

   (./data_raw/clothing-dataset-1)

5. Convierte los datos en un formato que puedas manipular fácilmente (e.g., csv) sin cambiar los datos en sí (Github, pero sujeto al punto 2).

   Los datos ya están en un formato manejable: - Las imágenes se encuentran en formato JPG. - Las etiquetas de las imágenes se encuentran en un archivo CSV.

6. Asegúrate de que la información sensible ha sido omitida (nombres, direcciones, etc; anonimízalos) (GitHub, pero sujeto al punto 2).

   No existe información sensible, por lo que este paso ha sido excluido.

7. Revisa qué tipo de datos tienes (¿son series de tiempo, datos geográficos, una combinación de diferentes tipos?).

   El tipo de datos son 10,000 imágenes de prendas de vestir en formato JPG divididos en dos conjuntos de 5000 donde el primer conjunto son las imágenes en resolución original y el segundo en una version comprimida.

8. Toma una muestra de prueba (test set), separarla y no la uses (GitHub, pero sujeto al punto 2).

Explorar datos

1. Crea una copia de los datos para su exploración. Si tienes un conjunto muy pesado, toma una muestra (aleatoria) de un tamaño manejable. En esta parte solo querrás hacerte una idea de qué es lo que tienes, y una muestra aleatoria de tamaño manejable asegura que las propiedades estadísticas de los datos son representativas del conjunto original.
2. Crea una notebook de Jupyter o un archivo de Quarto para la exploración. Es preferible una combinación de texto y código para documentar el proceso.
3. Estudia todos los atributos

   i. Nombres
   ii. Tipos de datos (categóricos/numéricos, enteros/puno flotante, restringidos o no restringidos, estructurados o no estructurados, etc).

iii. Porcentaje de datos perdidos.
iv. Ruido y su tipo (meramente estocástico, debido a error de redondeo, outliers posibles, o imposibles como pesos negativos, edades de 200 años, fechas imposibles, etc.)
v. Tipo de distribución de los datos.

4. Para algoritmos de supervisión, identifica la/las variables objetivo.
5. Visualiza los datos.
6. Estudia la correlación (u otras dependencias no lineales; por ejemplo esto) entre atributos.
7. Estudia cómo podrías resolver el problema manualmente.
8. Estudia las transformaciones que podrías aplicar (e.g., si una log-transformación vuelve normales los datos, sería mejor).
9. Identifica si obtener más datos podría ser útil. 10 .Documenta lo que has aprendido sobre los datos

Preparar los datos

> **Notas**
>
> Trabaja en copias de los datos (deja el original intacto). Escribir funciones para todas las transformaciones que apliques, de tal manera que: Puedas aplicarlo fácilmente la próxima vez que tengas datos frescos Apliques las mismas transformaciones en futuros proyectos Limpies y prepares el conjunto de prueba

1. Limpieza de datos Arreglar o remover outliers Rellenar datos perdidos (e.g., usando imputación múltiple, la media, o la mediana), o quitar las filas o columnas con NAs
2. Selección/ingeniería de características Remueve los atributos/variables que no proveen información para la tarea (opcional) Discretiza variables continuas, etc. Descompón características Añade transformaciones prometedoras (e.g., log-transformar, transformar con raíz cuadrada, etc., para volver normal)
3. Normaliza o estandariza las características E.g., Unit-based normalization. Esto es importante cuando tienes variables de diferentes escalas